

Electronic Companion to: Tutorial: Extracting Unstructured Text Using Large Language Models

FDA AdCom Meeting Transcript Evaluation

To understand the performance of our pipeline, we conduct a quantitative assessment of its accuracy whilst evaluating the costs of this extraction approach. We leverage the context of U.S. Food and Drug Administration (FDA) Advisory Committee (AdCom) meetings to conduct our analysis. FDA AdComs convene several times a year to deliberate on issues affecting public health, including guidance on medical best practices, new drug approvals, and post-market safety evaluations of approved therapies. The FDA AdCom context is ideal for benchmarking our approach for several reasons. First, verbatim meeting transcripts are publicly available and have previously been used for studying speech patterns in group decision-making (e.g., Urfalino et al. 2015, Markou and Chan 2026). Second, obtaining highly accurate textual data is crucial for downstream analysis, especially given the fundamental societal and policy implications stemming from AdCom deliberations. At the same time, reliance on a single institutional setting could raise concerns about context-specific tuning or overfitting of the extraction procedure. We therefore test our method’s generalizability to transcripts from meetings of the Federal Reserve’s Federal Open Market Committee in the last section.

Sample Selection We randomly select 10 verbatim meeting transcripts from Oncologic Drugs AdCom meetings between 1998-2016. Randomly selecting transcripts helps us test the robustness of our approach to evolving PDF formats over the years. For example, over this time period, the FDA changed its transcription service provider multiple times.

Table 1 gives an overview of the 10 selected FDA AdCom Oncology Drug meeting transcripts. They range from 132 to 326 pages in length. All of the selected transcripts included a text layer in their PDFs. Converting this text layer to tokens gives what we call the “Raw Tokens” count. The “Cleaned Tokens” column shows the token count after executing our vision-based text extraction Stage I. Note the large difference as compared to not using vision-based extraction, with an average reduction of 33.6%, a minimum reduction of 11.2%, and a maximum reduction of 52.2%. In other words, applying the vision-based stage provides a noticeably “cleaner” input to the data structuring stage. This increases the overall accuracy of the input to the downstream pipeline stages and simultaneously reducing subsequent token processing costs.

Extraction Approaches For our LLM-based extraction approach, we passed the entirety of the 10 PDFs through our pipeline. Specifically, we converted each page into PNG at default resolution. Each of these images was then converted to text using the Stage I vision-based LLM prompt and subsequently concatenated together into a single text file per document for the formatting Stage II. For both stages, we set the model temperature to 0. Due to the large size of the inputs (above 70,000 tokens in our sample), passing the entirety of the text is infeasible due to limitations in output length models such as `gpt-4.1-mini`. Even with larger possible output token limits, we saw decreased performance with larger chunk sizes. Therefore, we split the files into manageable 5,000 token “chunks” that cover approximately 20 raw PDF pages. For this evaluation, we used a fixed-token splitting method with additional logic to minimize splitting speaker’s statements.

Table 1 Summary of Selected FDA AdCom Oncology Drug Transcripts

Transcript #	Page Count	Raw Tokens	Cleaned Tokens	Reduction
1	326	82,729	73,497	11.2%
2	162	55,077	40,572	26.3%
3	166	52,454	36,562	30.3%
4	181	81,619	46,650	42.8%
5	244	72,390	51,189	29.3%
6	219	64,852	46,306	28.6%
7	177	52,212	38,624	26.0%
8	132	71,878	36,557	49.1%
9	132	78,802	37,668	52.2%
10	183	75,463	45,321	39.9%

Raw tokens correspond to extracts using direct extraction, whereas Cleaned
Tokens correspond to using our Vision extraction

We applied the previously described “UNKNOWN” speaker resolution from Listing 13 in the paper and secondary deterministic post-processing step at the end of Stage II. This step utilized regular expressions to remove parenthetical and bracketed text (e.g., [Laughter.] and (Applause.)). To assess the reliability and consistency of our extraction, we repeated the Stage II structuring five times for each document.

Performance Evaluation We evaluated the performance of this approach in terms of accuracy, speaker-attribution and cost. To assess accuracy, we measured the extent to which the models output aligns with the Ground Truth. We utilize the widely used Character Error Rate (CER) for document extraction (Neudecker et al. 2021), defined as:

$$CER = \frac{i + s + d}{n}$$

where n is the total number of characters in the reference text, and i , s , and d represent the number of character insertions, substitutions, and deletions required to transform the output into the Ground Truth. Because the number of edits can exceed the total number of characters, CER can theoretically exceed 1. For improved interpretability in our results, we report Character Accuracy (CAcc), where $CAcc = 1 - CER$. A value of 1 indicates a perfect match, whereas larger deviations from 1 indicate more errors.

Given the importance of correctly identifying the appropriate speaker, we also assess speaker-attribution performance. We treat this as a multi-class classification task and calculate the macro-F1 score for comparison. Specifically, we compare the reference speaker label with the model-predicted speaker label for each aligned sentence of speaker text. Sentences where the reference speaker is missing but the model predicts one are treated as extra predictions, while sentences where the reference speaker exists but the model gives no prediction are treated as missing predictions. For each speaker, we calculate precision and recall as:

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

where TP is the number of correctly attributed sentences for that speaker, FP is the number of sentences incorrectly attributed to that speaker, and FN is the number of sentences for that speaker that were either missed or incorrectly attributed to someone else. The speaker-level F_1 score is then calculated as:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

We then calculate macro- F_1 as the mean of the speaker-level F_1 scores, giving each speaker equal weight. A value of 1 indicates perfect speaker attribution, whereas a value of 0 indicates no correct speaker-level attribution under this scoring approach.

Finally, we examine the cost of this approach. We calculate the costs based on the `gpt-4.1-mini` pricing model. At the time of writing, OpenAI charges \$0.40 per million input tokens and \$1.60 per million output tokens for `gpt-4.1-mini`.

Table 2 Accuracy and Cost Results of Benchmarking Exercise

Transcript	Temperature = 0		Temperature = 1		Cost (\$)
	CAcc	F1-Score	CAcc	F1-Score	
1	0.99996	1.00000	0.99995	1.00000	0.59
2	0.99988	1.00000	0.99953	0.99879	0.30
3	0.99759	0.99919	0.99227	0.98321	0.29
4	1.00000	1.00000	0.99980	0.99987	0.34
5	0.99956	0.99830	0.99912	0.99748	0.42
6	0.99912	1.00000	0.99884	0.99583	0.38
7	0.99951	0.99912	0.99927	0.99912	0.31
8	0.99994	1.00000	0.99924	0.99630	0.26
9	0.99784	0.99935	0.99846	0.99957	0.27
10	0.99907	0.99673	0.99888	0.99398	0.34
Overall	0.99925	0.99927	0.99854	0.99642	3.50

Accuracy and F_1 -Score values for the LLM are averages over five passes per transcript.

LLM cost includes both vision-based processing and the average text extraction costs across 5 runs (see Figure 1).

Results Our LLM-based pipeline achieved excellent performance in terms of accuracy and speaker attribution while maintaining low cost. Across all 10 PDF transcripts, the recommended configuration (`temperature=0`) achieved an average CAcc of 0.99925 and an average macro- F_1 score for speaker attribution of 0.99927. For comparison, we also evaluated the pipeline with `temperature=1`. Although the resulting decrease in CAcc (0.99925 to 0.99854) and speaker attribution performance (macro- F_1 : 0.99927 to 0.99642) appears minor in absolute terms, they translate into substantially more transcription errors. Specifically, across all transcripts, the average number of character-level edits required to match the ground truth increased by 84% from 206 to 380. These results highlight the importance of using parameter settings that minimize variability in verbatim text extraction tasks.

Overall, it cost approximately \$3.50 to obtain an analysis-ready JSON file of all 10 meeting transcripts. Thus, the unit cost of our LLM approach was \$0.0018 per page. Figure 1 provides a detailed breakdown of the total LLM extraction costs reported in Table 2, into Stage I and II. The chart shows that the majority of the cost accrues from the vision-based LLM text extraction stage. Specifically, extracting text from images from the 10 transcripts cost \$2.61, whereas the text structuring process in Stage II cost a total of \$0.89.

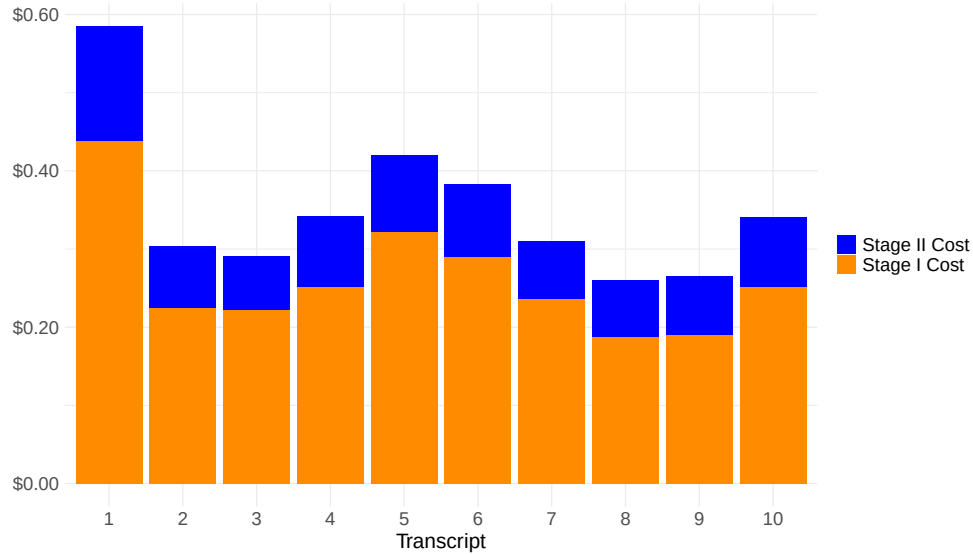


Figure 1 PDF Processing Costs

Federal Reserve Open Market Committee

To further evaluate the robustness of our pipeline, we apply the identical multi-stage process, without changing the prompts, to transcripts from the Federal Reserve Federal Open Market Committee (FOMC). As with FDA committees, the FOMC's influence on monetary policy makes its publicly available records a frequent subject of academic study (e.g., Hansen et al. 2018).

We randomly sample four FOMC meeting transcripts dating back to 1976 and extract 10 continuous pages from each to establish a Ground Truth dataset. These documents exhibit significant structural variation compared to FDA transcripts, including different speaker turn-taking conventions (e.g., periods versus colons following speaker names), diverse professional titles (e.g., "Vice Chairman"), and distinct page layouts. We execute the Stage I and II processes without any modification to generate structured JSON outputs.

Table 3 Results of FOMC Text Extraction

Transcript	# Pages	# Passes	Average CAcc
1	10	5	0.99834
2	10	5	0.98948
3	10	5	0.99758
4	10	5	0.99995
Overall			0.99634

Accuracy values are averages over five LLM passes per transcript PDF. Overall value reports the mean of transcript-level CAcc values.

Table 3 shows the results for this exercise. We find that our pipeline extracts textual data from the transcripts with high accuracy without any changes to the prompt. Our LLM-based approach produces fewer than 4 character errors

per 1,000 characters. We note that switching context can offer opportunities for tailored processing or additional post-processing, which our multi-stage approach can easily accommodate. These steps can help improve accuracy further. In our experience, many recurring issues are highly consistent across model runs; however, they may not be present in every document. For example, in one of the FOMC meeting transcripts, the term “M1” (as in ‘M1 money supply’) was frequently mis-extracted as “MI” (i.e., a ‘M’ followed by a lowercase letter ‘I’) during image transcription, whereas there were no issues in the remaining 3 samples with the same font type. We do not correct this issue within the accuracy calculation in Table 3 to demonstrate the generalizability of this approach, but suggest that there is still remaining low-hanging fruit for increased accuracy.

Impact of Temperature Control in Reasoning Models

With the introduction of reasoning models, several providers, including OpenAI and Anthropic, have reduced the degree of temperature control available to users. At the time of writing, our experiments indicate that OpenAI GPT-5 and newer models accepted only `temperature` settings of 0 and 1 (default) when `reasoning_effort=none`, with intermediate values resulting in API errors. To assess whether temperature still influences output variability, we executed the prompt from Listing 1 in the main article 50 times for each setting. With `temperature=0` we observed 16 unique outputs (ratio: 0.32), compared with 46 unique outputs (ratio: 0.92) for `temperature=1`. Consistent with the results reported in Table 2, these findings indicate that temperature remains an effective mechanism for reducing output variability, making it worthwhile to disable reasoning and set the temperature to zero when performing verbatim text extraction.

References

- Hansen S, McMahon M, Prat A (2018) Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics* 133(2):801–870.
- Markou P, Chan TH (2026) Incentivizing information exchange within groups: The role of voting protocols in us food and drug administration advisory committees. *Management Science* 72(7):5723–5743.
- Neudecker C, Baierer K, Gerber M, Clausner C, Antonacopoulos A, Pletschacher S (2021) A survey of ocr evaluation tools and metrics. *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, 13–18.
- Urfalino P, Costa P, Elster J (2015) *Secret-public voting in FDA advisory committees* (Cambridge University Press New York).